Patent Application of

Stephen J. Brown

for

Phenoscope and Phenobase

RELATED APPLICATION INFORMATION

This application is related to copending patent application 08/946,341 filed October 7, 1997 which is herein incorporated by reference.

FIELD OF THE INVENTION

This invention relates generally to the fields of genomics, bioinformatics, and drug development. More specifically, it relates to a database containing phenotypic and environmental data on groups of individuals for use in conjunction with gene sequences to identify disease-influencing genes and substances.

BACKGROUND OF THE INVENTION

The physical makeup of an individual is determined by his or her **genes**. Genes are comprised of **DNA**, which in turn consists of four **nucleotides** known as adenine(A), thymine(T), cytosine(C), and guanine(G). A particular series of nucleotides, ~~such as ATCGATCGATCG,~~ is known as a **gene sequence**. Each gene sequence codes for a **protein**. A defective or mutant gene sequence will not produce a working protein. The protein may not perform its purpose, the protein may carry out a different purpose than intended, too much protein may be made, too little protein may be made, or the protein may not be made at all. If the protein is essential to one or more functions of the body, disease will result.

Mutant gene sequences are either **inherited** or **acquired**. An inherited gene sequence is received from an individual's parents, while an acquired gene sequence results from an event in the individual's lifetime which changes the original gene sequence.

A classic example of an inherited mutant gene sequence is the sickle cell anemia gene. Sickle cell anemia is caused by the substitution of a single nucleotide (A to T) in the gene sequence of an individual. This single substitution results in the substitution of a single amino acid (glutamic acid to valine) in the resulting hemoglobin protein. The mutant hemoglobin protein produces crescent-shaped or sickled red blood cells in affected individuals, causing a decrease in the amount of oxygen that can be transported throughout the body. The lack of oxygen often results in kidney and heart failure, paralysis, and rheumatism, which are common symptoms of anemic individuals.

An example of an acquired mutant gene sequence is malignant melanoma, or skin cancer. Cancer results when normal cells in an individual's body either lose or gain certain functions, resulting in the unchecked growth of non-normal cells. These non-normal cells often form tumors and spread throughout the body, disrupting normal cell functions. A cancer such as malignant melanoma is caused when the original gene sequence in epidermal cells is changed or mutated by an environmental factor, such as UV radiation. Our cells contain repair mechanisms to fix such problems, but over time the gene sequences in epidermal cells acquire more and more mutations. Mutant proteins are then produced and cellular functions are disrupted. The individual then has skin cancer.

Although an individual's **environment** generally precipitates the development of cancer, many individuals have been found to have a predisposition to cancer. These individuals have gene sequences which are more likely to become mutated over a shorter

period of time. Examples of such gene sequences are the BRCA1 and BRCA2 genes. Women carrying these gene sequences have a higher probability of developing breast and ovarian cancer than women who carry normal gene sequences. Thus, although the affected women's original gene sequences may not be mutated, they are more likely to become mutated due to their sequence or location on a chromosome.

Another factor that should be considered when discussing genetic diseases is whether they are **monogenic** or **polygenic** in nature. Sickle cell anemia and cystic fibrosis are examples of monogenic diseases, as they are caused by a single gene sequence. Most types of cancer, asthma, and diabetes are examples of polygenic diseases, as they are caused by a variety of genes. Polygenic diseases are also more likely to be influenced by an individual's environment. Not surprisingly, polygenic diseases are more difficult to diagnose and treat. Thus, the use of gene sequences in developing new drugs is dependent the monogenic or polygenic nature of genetic diseases.

Typically, individuals with diseases caused by inherited or acquired gene sequences have only their symptoms treated. Diabetes patients receive insulin shots to regulate their blood glucose levels, asthma patients use inhalers to allow normal respiratory functions, and cancer patients undergo chemotherapy and radiation therapy to remove cancerous tumors. Although these treatments are often able to alleviate or eliminate the symptoms, they are unable to remove the genetic bases of the diseases.

The genetic bases of many diseases were discovered in the 1940's by scientists such as Beadle and Tatum, who discovered that each gene codes for a protein. Researchers then rationalized that study of the relevant gene sequences could lead to effective drug treatments for genetic diseases. The technology was inadequate, however, until the 1970-80's, when Boyer and Cohen

cloned DNA; Maxam, Gilbert, and Sanger figured out how to sequence DNA; and Mullis developed the polymerase chain reaction (PCR) technique to quickly amplify DNA sequences. Using genetics to find drug candidates soon became a practical option.

Before these techniques became available, the pharmaceutical industry's main method of finding new drugs was trial and error. Compounds that were found to mimic the body's natural compounds were tested *in vitro*, in animal models, and in clinical trials to see if they had a desirable effect in treating disease. This method is still used and has resulted in many well-known drugs, but it is expensive and time-consuming.

With the advent of improved genetic techniques, however, the pharmaceutical industry has begun concentrating on genetics as the most effective route to new drug discovery. Genomics companies can typically be classified into one of two groups.

The first group concentrates on **gene sequencing** in order to find both drug targets and drug candidates, usually in the form of proteins expressed by the gene sequences. Gene sequencing can either be in the form of random discovery, whereby genes are sequenced without regard to their functions, or in the form of targeted discovery, whereby a certain region of the genome which is tentatively associated with a disease is sequenced. In **random discovery gene sequencing**, potentially useful gene sequences are identified and assayed to determine if they can be used in drug development. One problem with random discovery gene sequencing is that the majority of the human genome contains introns, or gene sequences which do not code for proteins. One way to circumvent this problem is to sequence complementary DNA (cDNA) instead. cDNA is produced from messenger RNA (mRNA). mRNA, in turn, is transcribed from DNA and processed by certain enzymes which remove the introns. cDNA sequences thus code for un-interrupted proteins.

Targeted discovery gene sequencing is typically used with positional cloning, comparative gene expression, and functional cloning techniques, which are described in the next group.

The second group of genomics companies takes a more epidemiological approach by first researching families or groups of individuals having a similar disease, and then isolating the relevant genes. In this method, also known as **positional cloning**, blood samples are taken from the individuals and analyzed. The blood samples contain DNA, which is studied to identify certain regions of the genome which appear to be associated with the disease. Linking a region of the genome with a disease is known as **linkage analysis** or **genetic linkage mapping**. Once a region of the genome has been identified, it is sequenced via targeted discovery gene sequencing.

The second group of genomics companies also uses **comparative gene expression** to discover disease gene sequences. In comparative gene expression, mRNA from both healthy and diseased tissue is isolated. The mRNA is then used to produce cDNA, which is sequenced using targeted discovery gene sequencing. The gene sequences from both the healthy and diseased tissue are then compared. In addition, the identification of genes associated with disease can be made by studying the level of expression of genes in both the healthy and diseased tissue.

Another similar technique is **functional cloning**. Mutant or non-functional proteins in metabolic pathways are studied and identified. The proteins are sequenced using targeted discovery gene sequencing and these sequences are used to figure out the corresponding DNA gene sequences. Once the disease gene sequences have been identified, they can be used in drug development.

Genomics companies in the first group include **Incyte Pharmaceuticals** (Palo Alto, California). Incyte uses random

discovery gene sequencing to produce its LifeSeq™ and LifeSeq FL™ databases. These databases contain the sequences of hundreds of human genes. These databases are licensed to drug development companies who use the sequences to produce new drugs. Databases covering animals (ZooSeq™), plants (PhytoSeq™), and bacteria and fungi (PathoSeq™) are also available. Incyte has also developed bioinformatics software, which provides sequence analysis and data management for their databases. In addition, Incyte offers cDNA libraries of the gene sequences in their databases, which can be directly used in drug development.

Human Genome Sciences (Rockville, Maryland) also concentrates on random discovery gene sequencing, and has sequenced an estimated 90% of the 100,000 genes in the human body. In addition to collaborating with drug development companies who use their gene sequences, HGS also has its own drug discovery and development division. A number of therapeutic proteins which appear effective in animal models are under study.

Hyseq, Inc. (Sunnyvale, California) has its HyX Platform which is capable of processing and sequencing millions of blood and DNA samples. The HyX Platform includes DNA arrays of samples and probes, software-driven modules, industrial robots for screening DNA probes against DNA samples, and bioinformatic software to analyze the genetic information. Through the use of its HyX Platform, HyX believes it can carry out a variety of techniques, such as gene identification, gene expression level determination, gene interaction studies (for polygenic diseases), and genetic mapping.

Affymetrix, Inc. (Santa Clara, California) has a GeneChip system consisting of disposable DNA probe arrays containing gene sequences on a chip, instruments to process the probe arrays, and software to analyze and manage the genetic information in the probe. The GeneChip system thus allows pharmaceutical and

biotechnology companies to collect gene sequences and apply them to drug development.

On the other hand, the pharmaceutical industry has a number of genomics companies who first identify the genes which are likely to cause disease. After the genes are identified, they are sequenced and the gene sequences are used in drug development. Likewise, proteins implicated in disease can be identified and sequenced. The sequences can be used to discover the gene sequences, which are then used in drug development.

Myriad Genetics, Inc. (Salt Lake City, Utah) targets families with a history of genetic disease and collects their genetic material in order to identify hereditary disease-causing genes. Myriad is able to identify these genes by using positional cloning and protein interaction studies in combination with targeted discovery gene sequencing. Using these techniques, Myriad has been able to locate and identify eight disease-related gene sequences, including BRCA1 and BRCA2. These gene sequences are used by Myriad's pharmaceutical partners to develop new therapeutics.

Another genomics company which uses disease inheritance patterns together with gene sequencing is **Sequana** (La Jolla, California). Sequana uses DNA collection of individuals with inherited diseases, genotyping and linkage analysis, physical mapping, and gene sequencing to find disease gene sequences. Sequana also has a proprietary bioinformatics system which includes data mining tools to automatically sort and organize much of its data. Like Myriad, Sequana has a number of alliances with drug development companies which license Sequana's gene sequences.

Millennium Pharmaceuticals, Inc. (Cambridge, Massachusetts) employs a broader range of technologies than Myriad and Sequana. In addition to positional cloning and targeted discovery gene sequencing, Millennium uses a number of other non-genetic

techniques. cDNA libraries are prepared from mouse tissues and expressed using rapid expression of differential gene expression (RARE) technology.. Different patterns of cDNA gene expression allow researchers to identify possible disease targets. Millennium also uses functional cloning techniques in order to identify the gene sequences of interesting proteins. Once a potentially useful gene sequence has been identified, biological assays and bioinformatics are used as additional analyses.

Genome Therapeutics Corporation (Waltham, Massachusetts) uses a combination of positional cloning techniques and targeted discovery gene sequencing, as well as random discovery gene sequencing to isolate and identify disease gene sequences. In addition, Genome Therapeutics also has pathogen programs, which sequence pathogen genomes. As many non-genetic human diseases result from infection by pathogens, Genome Therapeutics hopes to eliminate pathogens by developing drugs and vaccines using the pathogens' genomes.

Gene Logic, Inc. (Columbia, Maryland) has an accelerated drug discovery system which emphasizes its **restriction enzyme analysis of differentially expressed sequences** (READS) technology. READS is similar in nature to comparative gene expression technology. In READS, normal and diseased tissues are compared in order to identify gene expression differences between the two. Genes which appear to be important in the diseased tissue are then analyzed. Restriction enzymes, which cut gene sequences at specific sites, are used to produce gene fragments. The gene fragments from the normal and diseased tissues will differ and can be compared. Gene Logic also has a Flow-thru Chip and genomic databases, which it licenses to drug development companies.

Progenitor (Columbus, Ohio) focuses on developmental biology. Growing cells and tissues are analyzed for their level of expression of certain genes. Study of growing cells and tissues

may help discover treatments for diseases characterized by abnormal cell growth, such as cancer and osteoporosis. Progenitor also uses bioinformatics, gene mapping, and gene sequencing to isolate, identify, and sequence relevant gene sequences.

OncorMed, Inc. (Gaithersburg, Maryland) has focused on the development of medical services using genetic information. Oncormed offers a number of tests for hereditary diseases such as breast and colon cancers and malignant melanoma. The medical services include measurements of replication error rates in tumors, molecular profiling of tumor suppresser genes, and gene sequencing. In addition, OncorMed has a genomics repository containing known cancer gene sequences.

U.S. Patent No. 5,642,936 issued to Evans and assigned to OncorMed describes a method for identifying human hereditary disease patterns. According to the method, data is collected on individuals having a history of disease within their families. Factors related to each disease are given weights, and the weights for each individual are summed. If the sum is above a certain predetermined threshold value, the individual is deemed to have a hereditary risk for the disease. Records from a number of individuals having a hereditary risk for a disease are collected to form a database.

The methods used by the above companies all focus on the genetic aspect of hereditary disease. Gene sequencing and positional cloning represent the two approaches generally taken. However, very little emphasis is put on the environmental aspect of hereditary disease. An individual's environment is defined as his or her physical surroundings, geographical location, diet, lifestyle, etc. For many diseases which are genetic in origin, such as most cancers, an individual's environment plays a large role in determining whether or not the individual eventually develops the disease. Some individuals who have disease gene

sequences develop diseases, while others who carry the exact same disease gene sequences do not. One purpose of collecting environmental data about individuals whose gene sequences are studied is to effectively rule out any non-genetic causes of disease. Another purpose is to discover if any individuals who are carrying disease gene sequences but who do not develop the disease have other compensatory gene sequences or factors which enable them to live disease-free.

To a certain extent, the second group of genomics companies do take into account a small amount of environmental data when they select individuals whose DNA they use for positional cloning analyses. The environmental data is usually in the form of a questionnaire or survey. However, the data is typically limited in scope to lifestyle questions, and is used only to help narrow the search for the specific disease gene in question.

In addition, most genomics companies are reluctant to share their data on individuals' with others, even those genomics companies which are studying the same gene sequences. As a result, each genomics company must gather its own data on individuals having a certain disease. For example, Sequana sent its own researcher to the island of Tristan de Cunha to study hereditary asthma, while Myriad is located in Salt Lake City to take advantage of the detailed family trees of the Mormons. For genomics companies searching for gene sequences, gathering environmental data on individuals is often an expensive, time-consuming, but necessary step. Genomics companies could potentially spend more of their time and money on actual disease gene isolation if they were able to obtain necessary environmental data from another source.

Another problem lies in the fact that when genomics companies do gather environmental data on the individuals whose gene sequences are studied, the environmental data represents only a small time frame of an individual's life. Few genomics

companies continually collect data over a long period of time, and as a result, are not able to definitively rule out certain environmental factors which may affect disease progression. In addition, such data collections are unlikely to provide leads for factors which may prohibit the formation of disease.

OBJECTS AND ADVANTAGES OF THE INVENTION

Accordingly, it is a primary object of the present invention to provide a system and method for creating a database of information about individuals' environments over a period of time. Another object of the present invention is to provide a database containing information about individuals' environments which can be used with existing genomics databases. A further object of the present invention is to provide a method of using environmental information about an individual in conjunction with the individual's genotype to find disease-influencing genes or substances. It is another object of the present invention to use the disease-influencing genes or substances to find drug candidates or drug targets.

SUMMARY OF THE INVENTION

These objects and advantages are attained by a system and method for identifying a disease-influencing gene or protein. The method includes the step of selecting individuals having a risk factor for a certain disease. Each of the individuals is provided with a remotely programmable apparatus having a user interface for communicating queries to the individuals and for receiving responses. Each apparatus also includes a communication device, such as a modem, for communicating with a server through a communication network.

Queries relating to the individuals' environment are entered into the server and transmitted from the server to each individual's remote apparatus. After the individuals' have responded to the queries, the responses are sent back to the server and organized into a database. Data mining software is

then used to distinguish the individuals into groups based on their environmental profiles. After a period of time, each group is then further divided into categories based on their disease progression. The genomes of all the individuals are then sequenced. Data mining techniques are used to find gene differences between the categories.

According to a second method of the invention, the individuals are first separated into groups according to their disease progressions. Data mining techniques are then used to further distinguish each group into categories based on the individuals' environmental profiles. The genomes of all the individuals are then sequenced, and data mining techniques are used to find gene differences between the categories.

A third embodiment of the invention provides a method for identifying disease-influencing substances. The method includes the step of selecting individuals having a risk factor for a certain disease. Each of the individuals is provided with a remotely programmable apparatus having a user interface for communicating queries to the individuals and for receiving responses. Each apparatus also includes a communication device, such as a modem, for communicating with a server through a communication network.

Queries relating to the individuals' environment are entered into the server and transmitted from the server to each individual's remote apparatus. After the individuals' have responded to the queries, the responses are sent back to the server and organized into a database. The genomes of all the individuals are then sequenced. The individuals are placed into groups based on their gene sequences. Each group is then separated into categories based on the individuals' disease progression. Data mining techniques are then used to find a disease-influencing substance between the categories of individuals by using the individuals environmental profiles.

The disease-influencing gene or substance isolated using these methods is preferably used to develop drug candidates or drug targets. Additionally, the isolation of the disease-influencing gene is preferably used to identify a corresponding disease-influencing protein, which can also be used to develop drug candidates or drug targets.

The present invention also provides a database and data processing system for storing and analyzing environmental information about individuals. The database and data processing system comprise a server for storing queries and the individuals' responses to the queries. The system also includes at least one remotely programmable apparatuses having a user interface for communicating queries to the individuals and for receiving the responses. Each apparatus also includes a communication device, such as a modem, for communicating with the server through a communication network.

The system also includes genotyping means in communication with the server for determining the individuals' gene sequences and a data mining software program accessible to the server for analyzing the individuals' gene sequences and environmental profiles. In particular, the data mining program includes: means for analyzing the responses in order to group the individuals having a similar behavioral and environmental profile, a similar disease progression, and a similar genotype; means for analyzing the responses in order to group the individuals having a similar disease progression; means for analyzing the responses in order to group the individuals having a similar genotype; and means for identifying a disease-influencing gene or substance. Alternatively, the database can be used with other genomics or bioinformatics databases and systems if the information is to be manipulated in different ways.

DESCRIPTION OF THE FIGURES

Fig. **1** is a block diagram of a networked system according to a preferred embodiment of the invention.

Fig. **2** is a block diagram illustrating the interaction of the components of the system of Fig. **1**.

Fig. **3** is a perspective view of a remotely programmable apparatus of the system of Fig. **1**.

Fig. **4** is a block diagram illustrating the components of the apparatus of Fig. **3**.

Fig. **5** is a script entry screen according to the preferred embodiment of the invention.

Fig. **6A** is a listing of a sample script program according to the preferred embodiment of the invention.

Fig. **6B** is a continuation of the listing of Fig. **6A**.

Fig. **7** is a script assignment screen according to the preferred embodiment of the invention.

Fig. **8** is a sample query appearing on a display of the apparatus of Fig. **3**.

Fig. **9** is a sample prompt appearing on the display of the apparatus of Fig. **3**.

Fig. **10** is a sample report displayed on a workstation of the system of Fig. **1**.

Fig. **11A** is a flow chart illustrating the steps included in a monitoring application executed by the server of Fig. **1** according to the preferred embodiment of the invention.

Fig. **11B** is a continuation of the flow chart of Fig. **11A**.

Fig. **12A** is a flow chart illustrating the steps included in the script program of Figs. **6A** - **6B**.

Fig. **12B** is a continuation of the flow chart of Fig. **12A**.

Fig. **13** is a sample data table of the present invention.

Fig. **14** is a sample completed data table of the present invention.

Fig. **15** is a flow chart illustrating a first method for identifying a gene according to the present invention.

Fig. **16** is a block diagram illustrating the method of Fig. **15**.

Fig. **17** is a flow chart illustrating a second method for identifying a gene according to the present invention.

Fig. **18** is a block diagram illustrating the method of Fig. **17**.

Fig. **19** is a flow chart illustrating a third method according to the present invention.

Fig. **20** is a block diagram illustrating the method of Fig. **19**.

DETAILED DESCRIPTION

The invention presents a system and method for creating a database containing environmental information about an individual to be used in conjunction with the individual's gene sequences to find new drug targets and drug candidates. In a preferred embodiment of the invention, remote monitors are used to collect the environmental information. It is to be understood that environmental information includes all non-genetic information about an individual, such as disease progression, diet, lifestyle, and geographical location.

A preferred embodiment of the invention is illustrated in Figs. **1 - 16**. Referring to Fig. **1**, a networked system includes a server **50** and a workstation **52** connected to server **50** through a communication network **58**. Server **50** is also connected to a patient profile database **54** which stores environmental information about the individuals. Server **50** is further connected to a genotyping system **56** which is capable of sequencing individuals' genomes. Patient profile database **54** and genotyping system **56** are connected to server **50** through communication network **58**.

Server **50** and patient profile database **54** are preferably world wide web servers. Server **50** and database **54** may comprise single stand-alone computers or multiple computers distributed throughout a network. Workstation **52** is preferably a personal computer, remote terminal, or web TV unit. Workstation **52**

functions as a remote interface for entering in server **50** messages and queries to be communicated to the individuals.

Genotyping system **56** can be a laboratory capable of sequencing individuals' genomes, a gene sequencing chip such as the GeneChip by Affymetrix, or any other suitable genotyping system. Genotyping system **56** should be capable of transmitting information about the individuals' genomes to server **50**. Communication network **58** connects workstation **52**, patient profile database **54**, and genotyping system **56** to server **50**. Communication network **58** can be any suitable communication network, such as a telephone cable, the Internet, or cellular or wireless communication. Such communication networks are well known in the art.

The system also includes remotely programmable apparatuses **60** for monitoring individuals. Preferably, each remote apparatus **60** is used to monitor a respective one of the individuals. Alternatively, a multi-user apparatus may be used to monitor a plurality of individuals. Each remote apparatus is designed to interact with an individual in accordance with script programs received from server **50**.

Each remote apparatus is in communication with server **50** through communication network **58**, which is preferably the Internet. Alternatively, each remote apparatus may be placed in communication with the server via telephone cable, cellular communication, wireless communication, etc. For clarity of illustration, only two remote apparatuses are shown in Fig. **1**. It is to be understood that the system may include any number of remote apparatuses for monitoring any number of individuals.

In the preferred embodiment, each individual to be monitored is also provided with a monitoring device **64**. Monitoring device **64** is designed to produce measurements of a physiological condition of the individual, record the measurements, and transmit the

measurements to the individual's remote apparatus **60** through a standard connection cable **62**. Examples of suitable monitoring devices include blood glucose meters, respiratory flow meters, blood pressure cuffs, electronic weight scales, and pulse rate monitors. Such monitoring devices are well known in the art.

The specific type of monitoring device provided to each individual is dependent upon the individual's disease. For example, diabetes patients are provided with blood glucose meters for measuring blood glucose concentrations, asthma patients are provided with respiratory flow meters for measuring peak flow rates, obesity patients are provided with weight scales, etc.

Fig. **2** shows server **50**, workstation **52**, and remote apparatus **60** in greater detail. Server **50** includes a database **66** for storing script programs **68**. The script programs **68** are executed by each remote apparatus **60** to communicate queries and messages to an individual, receive responses **70** to the queries, collect monitoring device measurements **72**, and transmit responses **70** and measurements **72** to server **50**. Database **66** is designed to store the responses **70** and measurements **72**. Database **66** further includes a look-up table **74**. Table **74** contains a list of the individuals to be monitored, and for each individual, a unique individual identification code and a respective pointer to script program **68** assigned to the individual. Each remote apparatus **60** is designed to execute the assigned script program which it receives from server **50**.

Figs. **3** - **4** show the structure of remote apparatus **60** according to the preferred embodiment. Referring to Fig. **3**, remote apparatus **60** includes a housing **90**. Housing **90** is preferably sufficiently compact to enable the remote apparatus to be hand-held and carried by an individual. Remote apparatus **60** also includes a user interface for communicating queries to the individual and for receiving responses to the queries.

In the preferred embodiment, the user interface includes a display **92** and four user input buttons **98A**, **98B**, **98C**, and **98D**. Display **92** displays queries and prompts to the individual, and is preferably a liquid crystal display (LCD). The user input buttons **98A**, **98B**, **98C**, and **98D** are for entering responses to the queries and prompts. The user input buttons are preferably momentary contact push buttons. Although the user interface of the preferred embodiment includes a display and input buttons, it will be apparent to one skilled in the art of electronic devices that any suitable user interface may be used in remote apparatus **60**. For example, the user input buttons may be replaced by switches, keys, a touch sensitive display screen, or any other data input device. Alternatively, the display and input buttons may be replaced by a speech synthesis/speech recognition interface.

Three monitoring device jacks **96A**, **96B**, and **96C** are located on a surface of housing **90**. Device jacks **96A**, **96B**, and **96C** are for connecting remote apparatus **60** to a number of monitoring devices, such as blood glucose meters, respiratory flow meters, or blood pressure cuffs, through standard connection cables (not shown). Remote apparatus **60** also includes a modem jack **94** for connecting remote apparatus **60** to a telephone jack through a standard connection cord (not shown). Remote apparatus **60** further includes a visual indicator, such as a light emitting diode (LED) **100**. LED **100** is for visually notifying the individual that he or she has unanswered queries stored in remote apparatus **60**.

Fig. **4** is a schematic block diagram illustrating the components of remote apparatus **60** in greater detail. Remote apparatus **60** includes a microprocessor **102** and a memory **108** connected to microprocessor **102**. Memory **108** is preferably a non-volatile memory, such as a serial EEPROM. Memory **108** stores script programs received from the server, measurements received from

monitoring device **64**, responses to queries, and the individual's unique identification code. Microprocessor **102** also includes built-in read only memory (ROM) which stores firmware for controlling the operation of remote apparatus **60**. The firmware includes a script interpreter used by microprocessor **102** to execute the script programs. The script interpreter interprets script commands which are executed by microprocessor **102**. Specific techniques for interpreting and executing script programs in this manner are well known in the art.

Microprocessor **102** is preferably connected to memory **108** using a standard two-wire I^2C interface. Microprocessor **102** is also connected to user input buttons **98A**, **98B**, **98C**, and **98D**, LED **100**, a clock **112**, and a display driver **110**. Clock **112** indicates the current date and time to microprocessor **102**. For clarity of illustration, clock **112** is shown as a separate component, but is preferably built into microprocessor **102**. Display driver **110** operates under the control of microprocessor **102** to display information on display **92**. Microprocessor **102** is preferably a PIC 16C65 processor which includes a universal asynchronous receiver transmitter (UART) **104**. UART **104** is for communicating with a modem **114** and a device interface **118**. A CMOS switch **116** under the control of microprocessor **102** alternately connects modem **114** and interface **118** to UART **116**.

Modem **114** is connected to a telephone jack **119** through modem jack **94**. Modem **114** is for exchanging data with the server through the communication network. The data includes script programs which are received from the server as well as responses to queries, device measurements, script identification codes, and the individual's unique identification code which modem **114** transmits to server **50**. Modem **114** is preferably a complete 28.8 K modem commercially available from Cermetek, although any suitable modem may be used.

Device interface **118** is connected to device jacks **96A**, **96B**, and **96C**. Device interface **118** is for interfacing with a number of monitoring devices, such as blood glucose meters, respiratory flow meters, blood pressure cuffs, weight scales, or pulse rate monitors, through device jacks **96A**, **96B**, and **96C**. Device interface **118** operates under the control of microprocessor **102** to collect measurements **72** from monitoring devices **64** and to output measurements **72** to microprocessor **102** for storage in memory **108**. In the preferred embodiment, interface **118** is a standard RS232 interface. For simplicity of illustration, only one device interface **118** is shown in Fig. **4**. However, in alternative embodiments, remote apparatus **60** may include multiple device interfaces to accommodate monitoring devices which have different connection standards.

Referring again to Fig. **2**, server **50** includes a monitoring application **76**. Monitoring application **76** is a controlling software application executed by server **50** to perform the various functions described below. Monitoring application **76** includes a script generator **78**, a script assignor **80**, and a report generator **82**. Script generator **78** is designed to generate script programs **68** from script information entered through workstation **52**. The script information is entered through a script entry screen **84**. In the preferred embodiment, script entry screen **84** is implemented as a web page on the server **50**. Workstation **52** includes a web browser for accessing the web page to enter the script information.

Fig. **5** illustrates script entry screen **84** as it appears on workstation **52**. Script entry screen **84** includes a script name field **120** for specifying the name of script program to be generated. Screen **84** also includes entry fields **122** for entering a set of queries to be answered by an individual. Each entry field **122** has corresponding response choice fields **124** for entering response choices for the query. Screen **84** further includes check boxes **126** for selecting a desired monitoring

device type from which to collect measurements, such as a blood glucose meter, respiratory flow meter, or blood pressure cuff.

Screen **84** additionally includes a connection time field **128** for specifying a prescribed connection time at which each remote apparatus executing the script program is to establish a subsequent communication link to the server. The connection time is preferably selected to be the time at which communication rates are the lowest, such as 3:00 AM. Screen **84** also includes a CREATE SCRIPT button **130** for instructing the script generator to generate a script program from the information entered in screen **84**. Screen **84** further includes a CANCEL button **132** for canceling the information entered.

In the preferred embodiment, each script program created by the script generator **82** conforms to the standard file format used on UNIX systems. In the standard file format, each command is listed in the upper case and followed by a colon. Every line in the script program is terminated by a linefeed character {LF}, and only one command is placed on each line. The last character in the script program is a UNIX end of file character {EOF}. **TABLE 1** shows an exemplary listing of script commands used in the preferred embodiment of the invention.

TABLE 1 - SCRIPT COMMANDS

Command	Description
CLS: {LF}	Clear the display.
ZAP: {LF}	Erase from memory the last set of query responses recorded.
LED: b{LF}	Turn the LED on or off, where b is a binary digit of 0 or 1. An argument of 1 turns on the LED, and an argument of 0 turns off the LED.
DISPLAY: {chars}{LF}	Display the text following the DISPLAY command.
INPUT: mmmm{LF}	Record a button press. The m's represent a button mask pattern for each of the four input buttons. Each m contains an "X" for disallowed buttons or an "O" for allowed buttons. For example, INPUT: OXOX{LF} allows the user to press either button #1 or #3.
WAIT: {LF}	Wait for any one button to be pressed, then continue executing the script program.
COLLECT: device{LF}	Collect measurements from the monitoring device specified in the COLLECT command. The user is preferably prompted to connect the specified monitoring device to the apparatus and press a button to continue.
NUMBER: aaaa{LF}	Assign a script identification code to the script program. The script identification code from the most recently executed NUMBER statement is subsequently transmitted to the server along with the query responses and device measurements. The script identification code identifies to the server which script program was most recently executed by the remote apparatus.
DELAY: t {LF}	Wait until time t specified in the DELAY command, usually the prescribed connection time.
CONNECT: {LF}	Perform a connection routine to establish a communication link to the server, transmit the patient identification code, query responses, device measurements, and script identification code to the server, and receive and store a new script program. When the server instructs the apparatus to disconnect, the script interpreter is restarted, allowing the new script program to execute.

The script commands illustrated in **TABLE 1** are representative of the preferred embodiment and are not intended to limit the scope of the invention. After consideration of the ensuing description, it will be apparent to one skilled in the art many other suitable scripting languages and sets of script commands may be used to implement the system and method of the invention.

Script generator **78** preferably stores a script program template which it uses to create each script program. To generate a script program, script generator **78** inserts into the template the script information entered in script entry screen **84**. For example, Figs. **6A - 6B** illustrate a sample script program created by the script generator from the script information shown in Fig. **5**.

The script program includes display commands to display the queries and response choices entered in fields **122** and **124**, respectively. The script program also includes input commands to receive responses to the queries. The script program further includes a collect command to collect device measurements from the monitoring device specified in check boxes **126**. The script program also includes commands to establish a subsequent communication link to the server at the connection time specified in field **128**. The steps included in the sample script program are also shown in the flow chart of Figs. **12A - 12B** and will be discussed in the operation section below.

Referring again to Fig. **2**, script assignor **80** is for assigning the script programs **68** to the individuals. The script programs are assigned in accordance with script assignment information entered through workstation **52**. The script assignment information is entered through a script assignment screen **86**, which is preferably implemented as a web page on server **50**.

Fig. **7** shows a sample script assignment screen **86** as it appears on the workstation. Screen **86** includes check boxes **134** for

selecting the script program to be assigned and check boxes **136** for selecting the individuals to whom the script program is to be assigned. Screen **86** also includes an ASSIGN SCRIPT button **140** for entering the assignments. When button **140** is pressed, the script assignor creates and stores for each individual selected in check boxes **136** a respective pointer to the script program selected in check boxes **134**. Each pointer is stored in the look-up table **74** of database **66**. Screen **86** further includes an ADD SCRIPT button **138** for accessing the script entry screen and a DELETE SCRIPT button **142** for a deleting script program.

Referring again to Fig. **2**, report generator **82** is designed to generate a report **88** from the responses **70** and device measurements **72** received in server **50**. Report **88** is displayed on workstation **52**. Fig. **10** shows a sample patient report **88** produced by report generator **82** for a selected individual. Report **88** includes a graph **146** of the device measurements received from the individual, as well as a listing of the query responses received from the individual. Specific techniques for writing a report generator program to display data in this manner are well known in the software art.

The operation of the preferred embodiment is illustrated in Figs. **1 - 12**. Fig. **11A** is a flow chart illustrating steps included in the monitoring application executed by server **50**. Fig. **11B** is a continuation of the flow chart of Fig. **11A**. In step **202**, the server determines if new script information has been entered through script entry screen **84**. If new script information has not been entered, the server proceeds to step **206**. If new script information has been entered, the server proceeds to step **204**.

As shown in Fig. **5**, the script information includes a set of queries, and for each of the queries, corresponding responses choices. The script information also includes a selected monitoring device type from which to collect measurements. The

script information further includes a prescribed connection time for each remote apparatus to establish a subsequent communication link to the server. The script information is generally entered in the server by a healthcare provider, such as the individuals' physician or case manager. Of course, any person desiring to communicate with the individual may also be granted access to the server to create and assign script programs. Further, it is to be understood that the system may include any number of workstations for entering script generation and script assignment information into the server.

In step **204**, script generator **78** generates a script program from the information entered in screen **84**. The script program is stored in database **66**. Steps **202** and **204** are preferably repeated to generate multiple script programs, e.g. a script program for diabetes patients, a script program for asthma patients, etc. Each script program corresponds to a respective one of the sets of queries entered through script entry screen **84**. Following step **204**, the server proceeds to step **206**.

In step **206**, the server determines if new script assignment information has been entered through script assignment screen **86**. If new script assignment information has not been entered, the server proceeds to step **210**. If new script assignment information has been entered, the server proceeds to step **208**. As shown in Fig. **7**, script programs are assigned to each individual by selecting a script program through check boxes **134**, selecting the individuals to whom selected the script program is to be assigned through check boxes **136**, and pressing the ASSIGN SCRIPT button **140**. When button **140** is pressed, script assignor **86** creates for each individual selected in check boxes **136** a respective pointer to the script program selected in check boxes **134**. In step **208**, each pointer is stored in look-up table **74** of database **66**. Following step **208**, the server proceeds to step **210**.

In step **210**, the server determines if any of the remote apparatuses are remotely connected to the server. Each individual to be monitored is preferably provided with his or her own remote apparatus which has the individual's unique identification code stored therein. Each individual is thus uniquely associated with a respective one of the remote apparatuses. If none of remote apparatuses are connected, the server proceeds to step **220**.

If a remote apparatus is connected, the server receives from the apparatus the individual's unique identification code in step **212**. In step **214**, the server receives from the apparatus the query responses, device measurements, and script identification code recorded during execution of a previously assigned script program. The script identification code identifies to the server which script program was executed by the remote apparatus to record the query responses and device measurements. The responses, device measurements, and script identification code are stored in database **66**.

In step **216**, the server uses the individual's unique identification code to retrieve from look-up table **74** the pointer to the script program assigned to the individual. The server then retrieves the assigned script program from the database **66**. In step **218**, the server transmits the assigned script program to the individual's remote apparatus through the communication network **58**. Following step **218**, the server proceeds to step **220**.

In step **220**, the server determines if a report request has been received from workstation **52**. If no report request has been received, the server returns to step **202**. If a report request has been received for a selected individual, the server retrieves from database **66** the query responses and measurements last received from the individual, step **222**. In step **224**, the server generates and displays the report **88** on workstation **52**.

As shown in Fig. **10**, the report includes the query responses and device measurements last received from the individual. Following step **224**, the server returns to step **202**.

Figs. **12A - 12B** illustrate the steps included in a sample script program executed by the remote apparatus. Before the script program is received, the remote apparatus is initially programmed with the individual's unique identification code and the script interpreter used by microprocessor **102** to execute script programs. The initial programming may be achieved during manufacture or during an initial connection to the server. Following initial programming, the remote apparatus receives from the server the script program assigned to the individual associated with the apparatus. The script program is received by modem **114** through a first communication link to the server and stored in memory **108**.

In step **302**, microprocessor **102** assigns a script identification code to the script program and stores the script identification code in memory **108**. The script identification code is subsequently transmitted to the server along with query responses and device measurements to identify to the server which script program was most recently executed by the remote apparatus. In step **304**, microprocessor **102** lights LED **100** to notify the individual that he or she has unanswered queries stored in the remote apparatus. LED **100** preferably remains lit until the queries are answered by the individual. In step **306**, microprocessor **102** erases from memory **108** the last set of query responses recorded.

In step **308**, microprocessor **102** prompts the individual by displaying on display **92** "ANSWER QUERIES NOW? PRESS ANY BUTTON TO START". In step **310**, microprocessor **102** waits until a reply to the prompt is received from the individual . When a reply is received, microprocessor **102** proceeds to step **312**. In step **312**, microprocessor **102** executes successive display and input

commands to display the queries and response choices on display **92** and to receive responses to the queries.

Fig. **8** illustrate a sample query and its corresponding response choices as they appear on display **92**. The response choices are preferably positioned on display **92** such that each response choice is located proximate a respective one of the user input buttons **98A**, **98B**, **98C**, and **98D**. In the preferred embodiment, each response choice is displayed immediately above a respective user input button. The individual presses the button corresponding to his or her response, and microprocessor **102** stores the response in memory **108**.

In steps **314** to **318**, microprocessor **102** executes commands to collect device measurements from a selected monitoring device specified in the script program. In step **314**, microprocessor **102** prompts the individual to connect the selected device to one of the device jacks **96A**, **96B**, or **96C**. A sample prompt is shown in Fig. **9**. In step **316**, microprocessor **102** waits until a reply to the prompt is received from the individual. When a reply is received, microprocessor **102** proceeds to step **318**. Microprocessor **102** also connects UART **104** to device interface **118** through CMOS switch **116**. In step **318**, microprocessor **102** collects device measurements from the selected device through device interface **118**. The device measurements are stored in memory **108**.

In step **320**, microprocessor **102** prompts the individual to connect remote apparatus **60** to telephone jack **119** so that the apparatus may connect to the server at the prescribed connection time. In step **322**, microprocessor **102** waits until a reply to the prompt is received from the individual. When a reply is received, microprocessor **102** turns off LED **100** in step **324**. In step **326**, microprocessor **102** waits until it is time to connect to the server. Microprocessor **102** compares the connection time specified in the script program to the current time output by

clock **112**. When it is time to connect, microprocessor **102** connects UART **104** to modem **114** through CMOS switch **116**.

In step **328**, microprocessor **102** establishes a subsequent communication link between remote apparatus **60** and server **50** through modem **114** and communication network **58**. If the connection fails for any reason, microprocessor **102** repeats step **328** to get a successful connection. In step **330**, microprocessor **102** transmits the query responses, device measurements, script identification code, and the individual's unique identification code stored in memory **108** to the server. In step **332**, microprocessor **102** receives through modem **114** a newly assigned script program from the server. The new script program is stored in memory **108** for subsequent execution by microprocessor **102**. Following step **332**, the script program ends.

After the individual's information has been collected via remote apparatus **60** and the script programs, the data is mined to distinguish patterns. Data mining programs are well known in the art and can be easily adapted to this system. In the preferred embodiment, the data mining program includes a data table **150**, as shown in Fig. **13**. Data table **150** is stored on the server and has an individual identification number field **151**, name fields **152**, value fields **154** corresponding to the name fields, and explanation fields **156** corresponding to the name fields and value fields. The data type is entered into name fields **152**, the possible numerical values corresponding to the data type are entered into value fields **154**, and brief explanations of the data types and corresponding values are entered into explanation fields **156**.

The individuals' device measurements and responses to the queries are entered into data table **150** in the form of numerical values in value fields **154**. The individual's identification number is entered into individual identification number field **151**. An example of data table **150** in which the individuals'

information has been entered is shown in Fig. **14**. Once data table **150** contains all the necessary information, the data mining program then compares the information.

Fig. **15** is a flowchart illustrating a first method of the present invention carried out by the server using the data mining techniques described above. In step **400**, individuals having a risk factor for a disease are selected. In step **402**, these individuals are queried about their behavior and environment using the script programs and remote apparatuses previously described. The responses to the queries and any device measurements are received and stored by the server. Collection of the responses and device measurements can occur over any period of time, thus allowing for more accurate data.

After the server receives the responses and measurements, a database comprising the individuals' behavioral and environmental profiles is created in step **404**. In step **406**, data mining techniques are used to group individuals having similar behavioral and environmental profiles. In step **408**, the server determines if it is necessary to further group the individuals in order to produce smaller groups. Steps **406** and **408** can be repeated as often as necessary.

In step **410**, each group of individuals is categorized using data mining techniques. The individuals are categorized according to their disease progressions. For example, a group of individuals can be categorized into those that have a severe disease phenotype, those that have a moderate disease phenotype, and those that have a mild disease phenotype. In step **412**, the server determines if it is necessary to further categorize the individuals. Steps **410** and **412** can be repeated as often as necessary.

In step **414**, the genomes of all the individuals are sequenced by genotyping system **56**. The genotypes of all the individuals are

transmitted to server **50**. In step **416**, data mining techniques are used to compare the genotypes of the individuals between the categories. For example, if those individuals who have a severe disease phenotype and are overweight have a certain gene sequence, while those individuals who have a mild disease phenotype and are overweight do not, it is likely the gene sequence is responsible for the severe disease phenotype. If a gene sequence is found, it is further identified in step **418**. Methods of isolating and identifying gene sequences are well known in the field.

Fig. **16** is a block diagram illustrating an example of the first method of the present invention as described in Fig. **15**. First individuals having a risk factor for a certain disease, such as non-insulin dependent diabetes mellitus (NIDDM), are selected, as indicated at block **422**. Behavioral and environmental information from each individual is collected using the script programs and remote apparatuses. Using data mining techniques, the individuals are then grouped into overweight individuals **424** and non-overweight individuals **426**. Using data mining techniques, the individuals are then categorized into overweight individuals having severe NIDDM **428**, overweight individuals having mild NIDDM **430**, non-overweight individuals having mild NIDDM **432**, and non-overweight individuals having severe NIDDM **434**.

The individuals' genotype information is then taken, as indicated at block **436**, to determine the individuals' gene sequences. For example, overweight individuals with severe NIDDM have gene sequence A, overweight individuals with mild NIDDM have gene sequence B, non-overweight individuals with mild NIDDM have gene sequence B, and non-overweight individuals with severe NIDDM have gene sequence A. Data mining techniques are then used to analyze the information and come to a conclusion. In this example, data mining would conclude that the severe

NIDDM phenotype is likely related to gene sequence A, not the individual's weight.

Fig. **17** shows a flowchart illustrating a second method of the present invention carried out by the server using the data mining techniques described above. In step **500**, individuals having a risk factor for a disease are selected. In step **502**, these individuals are queried about their behavior and environment using the script programs and remote apparatuses previously described. The responses to the queries and any device measurements are received and stored by the server.

After the server receives the responses and measurements from the remote apparatuses, a database comprising the individuals' behavioral and environmental profiles is created in step **504**. In step **506**, data mining techniques are used to group together individuals having similar disease progressions. For example, a group of individuals can be grouped into those that have a severe disease phenotype, those that have a moderate disease phenotype, and those that have a mild disease phenotype. In step **508**, the server determines if it is necessary to further group the individuals in order to produce smaller groups. Steps **506** and **508** can be repeated as often as necessary.

In step **510**, each group of individuals created in steps **506** and **508** is categorized using data mining techniques according to the behavioral and environmental profiles of the individuals. In step **512**, the server determines if it is necessary to further group the individuals in order to produce smaller groups. Steps **510** and **512** can be repeated as often as necessary.

In step **514**, the genomes of all the individuals are sequenced by genotyping system **56**. The genotypes of all the individuals are transmitted to the server. In step **516**, data mining techniques are used to compare the genotypes of the individuals between the categories. For example, if those individuals who have a severe

disease phenotype and are overweight have a certain gene sequence, while those individuals who have a mild disease and are also overweight phenotype do not, it is likely the gene sequence, not weight, is responsible for the severe disease phenotype. If a gene sequence is found, it is further identified in step **518**. Specific techniques of isolating and identifying gene sequences are well known in the field.

Fig. **18** is a block diagram illustrating an example of the second method of the present invention as described in Fig. **17**. First individuals having a risk factor for a certain disease, such as NIDDM, are chosen, as indicated at block **522**. Behavioral and environmental information from each individual is collected using the remote apparatuses and script programs. Using data mining techniques, the individuals are then grouped into those exhibiting severe NIDDM **524** and those exhibiting mild NIDDM **526**. Using data mining techniques, the individuals are then categorized into overweight individuals having severe NIDDM **528**, non-overweight individuals having severe NIDDM **530**, non-overweight individuals having mild NIDDM **532**, and overweight individuals having mild NIDDM **534**.

The individuals' genotype information is then taken, as indicated at block **536**, to determine the individuals' gene sequences. For example, individuals with severe NIDDM who are overweight have gene sequence A, individuals with severe NIDDM who are non-overweight have gene sequence A, individuals with mild NIDDM who are non-overweight have gene sequence B, and individuals with severe NIDDM who are overweight have gene sequence B. Data mining techniques are then used to analyze the information and come to a conclusion. In this example, data mining would conclude that the severe NIDDM phenotype is likely related to gene sequence A, not the individual's weight.

Fig. **19** shows a flowchart illustrating a preferred method carried out by server **50** to identify a disease-identifying

substance. In step **600**, individuals having a risk factor for a disease are selected. In step **602**, these individuals are queried about their behavior and environment using the script programs and remote apparatuses previously described. The responses to the queries and any device measurements are received and stored by the server.

After the server receives the responses and measurements from the remote apparatuses, a database comprising the individuals' behavioral and environmental profiles is created in step **604**. In step **606**, the genomes of all the individuals are sequenced, and the genotypes of all the individuals are transmitted to the server. In step **608**, individuals having the same or close genotypes are grouped together. In step **610**, data mining techniques are used to categorize together individuals having similar disease progressions. In step **612**, the server determines if it is necessary to further categorize the individuals in order to produce smaller groups. Steps **610** and **612** can be repeated as often as necessary.

In step **614**, data mining techniques are used to find a disease-influencing substance between the categories of individuals by using the individuals behavioral and environmental profiles. For example, if those individuals who have a severe disease phenotype are overweight, while those individuals who have a mild disease phenotype are not, it is likely weight is responsible for the severe disease phenotype. If such a disease-influencing substance is found, it is identified in step **618**. If no disease-influencing substance is found, the process is preferably repeated.

Fig. **20** is a block diagram illustrating an example of the method described in Fig. **19**. First, individuals having a risk factor for a certain disease, such as NIDDM, are chosen, as indicated at block **620**. Behavioral and environmental information from each individual is collected using the remote apparatuses and

script programs. The individuals' genotype information is then taken, as indicated at block **622**, to determine the individuals' gene sequences. The individuals are then grouped according to their gene sequences. For example, one group may have gene sequence A, as indicated at block **624**, while another group may have gene sequence B, as indicated at block **626**. Using data mining techniques, the individuals are then categorized into individuals with gene sequence A having severe NIDDM **628**, individuals with gene sequence A having mild NIDDM **630**, individuals with gene sequence B having mild NIDDM **632**, and individuals with gene sequence B having severe NIDDM **634**.

Data mining techniques are further used to analyze the categories of individuals and their behavioral and environmental profiles. For example, overweight individuals **638** with severe NIDDM have gene sequence A, non-overweight individuals **640** with mild NIDDM have gene sequence A, overweight individuals **642** with mild NIDDM have gene sequence B, and non-overweight individuals **644** with severe NIDDM have gene sequence B. Data mining techniques are then used to analyze the information and come to a conclusion. In this example, data mining would conclude that the severe NIDDM phenotype is likely related to gene sequence A, not the individual's weight.

SUMMARY, RAMIFICATIONS, AND SCOPE

Although the above description contains many specificities, these should not be construed as limitations on the scope of the invention but merely as illustrations of some of the presently preferred embodiments. Many other embodiments of the invention are possible. For example, the scripting language and script commands shown are representative of the preferred embodiment. It will be apparent to one skilled in the art that many other scripting languages and specific script commands may be used to implement the invention.

Moreover, the invention is not limited to the specific applications described. The system and method of the invention have many other applications. For example, pharmaceutical manufacturers may apply the system in clinical trials to analyze new drug data.

Therefore, the scope of the invention should be determined by the appended claims and their legal equivalents.